

## Practice of Epidemiology

### Effects of Adjusting for Instrumental Variables on Bias and Precision of Effect Estimates

Jessica A. Myers\*, Jeremy A. Rassen, Joshua J. Gagne, Krista F. Huybrechts,  
Sebastian Schneeweiss, Kenneth J. Rothman, Marshall M. Joffe, and Robert J. Glynn

\* Correspondence to Dr. Jessica A. Myers, Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, 1620 Tremont Street, Suite 3030, Boston, MA 02120 (e-mail: jmyers6@partners.org).

Initially submitted February 28, 2011; accepted for publication June 8, 2011.

Recent theoretical studies have shown that conditioning on an instrumental variable (IV), a variable that is associated with exposure but not associated with outcome except through exposure, can increase both bias and variance of exposure effect estimates. Although these findings have obvious implications in cases of known IVs, their meaning remains unclear in the more common scenario where investigators are uncertain whether a measured covariate meets the criteria for an IV or rather a confounder. The authors present results from two simulation studies designed to provide insight into the problem of conditioning on potential IVs in routine epidemiologic practice. The simulations explored the effects of conditioning on IVs, near-IVs (predictors of exposure that are weakly associated with outcome), and confounders on the bias and variance of a binary exposure effect estimate. The results indicate that effect estimates which are conditional on a perfect IV or near-IV may have larger bias and variance than the unconditional estimate. However, in most scenarios considered, the increases in error due to conditioning were small compared with the total estimation error. In these cases, minimizing unmeasured confounding should be the priority when selecting variables for adjustment, even at the risk of conditioning on IVs.

bias (epidemiology); confounding factors (epidemiology); epidemiologic methods; instrumental variable; precision; simulation; variable selection

Abbreviations: IV, instrumental variable; RD, risk difference; RR, risk ratio.

**Editor's note:** An invited commentary on this article appears on page 1223, and the authors' response appears on page 1228.

In studies of exposure effect, measured and unmeasured factors that are associated with both exposure and outcome may confound the targeted causal effect. Estimating the exposure effect conditional on all confounding factors yields consistent estimates (1–4), so choosing which variables to use for adjustment in studies with many measured covariates is an important step for ensuring the validity of effect estimates. In an attempt to mimic a randomized trial, some authors have argued that all measured preexposure covariates should be balanced between exposure groups (5–8). This

strategy is equivalent to selecting all predictors of exposure, a common practice when confounder adjustment is carried out via the propensity score (9). Other authors have argued against this practice on the grounds that adjusting for some types of covariates may increase rather than decrease bias (10–12).

In particular, recent literature has questioned whether instrumental variables (IVs) (or instruments) should be conditioned upon in effect estimation. IVs are variables that are associated with exposure but are not associated with outcome, except through their effect on exposure. IVs may be used to obtain an unbiased estimate of exposure effect in the presence of unmeasured confounding via the class of IV methods. (See recent reviews (13–16) for precise definitions of IVs and IV methods.) Because IVs are, by definition, predictors of exposure, any confounder selection strategy

that is based on selecting the predictors of exposure will be likely to include IVs.

Rubin (17) suggested that including a variable that is unrelated to outcome in the propensity score may reduce estimation efficiency, and that result was confirmed in simulation studies (18, 19). Theoretical results presented by Hahn (20) and White and Lu (21) show that selecting confounders to maximize independent variation in the exposure will result in more efficient estimators. In addition, theoretical analyses have established that including IVs in the set of conditioning variables can increase unmeasured confounding bias (22–24), and empirical examples presented by Bhattacharya and Vogt (22) and Patrick et al. (25) found that including a “known” instrument in the propensity score model resulted in an estimate which was farther from the assumed truth than that obtained from the model that did not include the IV.

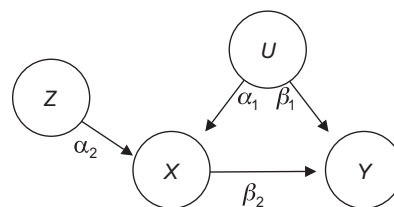
Despite the evident drawbacks of conditioning on an IV, the implication of these results for epidemiologic practice remains unclear. True instruments are difficult to identify and cannot be verified empirically (15). For example, in a series of commentaries on a paper by Stukel et al. (26), authors debated whether or not the assumed IV, regional cardiac catheterization rate, was more likely to be a confounder of the association between invasive cardiac management and survival of acute myocardial infarction and should therefore be adjusted for (27–30). Moreover, in the presence of unmeasured confounding, an IV may look mistakenly like a confounder, since it may be associated with exposure and associated with outcome conditional on exposure. Finally, the available theoretical studies of this issue provide results under linear models and do not indicate the magnitude of the increases in bias and variance for other models. Any increase in bias due to unnecessary conditioning must be weighed against the danger of excluding real confounders from the conditioning set—an issue that is particularly troubling in secondary analyses of electronic health-care data that often rely on adjusting for hundreds of confounding covariates (31, 32).

Our objective in the current analysis was to explore the magnitude of the effects on bias and variance of conditioning on an IV in a range of common epidemiologic studies of a binary exposure. Here we expand on the theoretical analyses by providing quantitative results under a range of common linear models and by further providing results under multiplicative models. We focus on the case where an IV may exist in the set of measured variables but it is uncertain to investigators. We present results from a Monte Carlo simulation study that considers true instruments, variables with no direct effect on unobserved confounding factors or outcome, and “near-instruments,” variables that are weakly associated with the unmeasured confounder. We also explore effects under varying assumptions about the strength of the IV association with exposure and the magnitude of the unmeasured confounding.

## MATERIALS AND METHODS

### Review of the theory

We refer to  $X$  as the exposure of interest and  $Y$  as an outcome that may be caused by  $X$ . We assume that there exists an unobserved factor,  $U$ , that confounds the association



**Figure 1.** Causal diagram showing an unmeasured confounder,  $U$ , and an instrumental variable,  $Z$ , of the exposure-outcome pair ( $X$ ,  $Y$ ).

between  $X$  and  $Y$  and a measured covariate,  $Z$ . If  $Z$  satisfies the criteria for an IV for the exposure-outcome pair ( $X$ ,  $Y$ ), then there is no association between  $Z$  and  $Y$ , except through  $X$ , as shown in Figure 1. We may think of this graph as representing residual associations after controlling for a vector of measured confounders. In addition,  $U$  may represent a constellation of many unobserved confounders, and  $Z$  may represent the combined effect of multiple instruments. The true exposure effect and target of estimation is  $\beta_2$ . The parameter  $\alpha_2$  controls the strength of the IV association with exposure. The magnitude of confounding is dependent on both  $\alpha_1$  and  $\beta_1$ .

We want to compare the bias of the crude, unadjusted estimator of exposure effect (given by the coefficient of the regression of  $Y$  on  $X$ ) with the bias of the estimator for exposure effect that conditions on  $Z$  (given by the regression coefficient on  $X$  in the regression of  $Y$  on  $X$  and  $Z$ ). We follow the example of Pearl (24) and assume a linear structural equation framework among zero-mean, unit-variance variables. Under these assumptions, the crude association between  $X$  and  $Y$  is given by

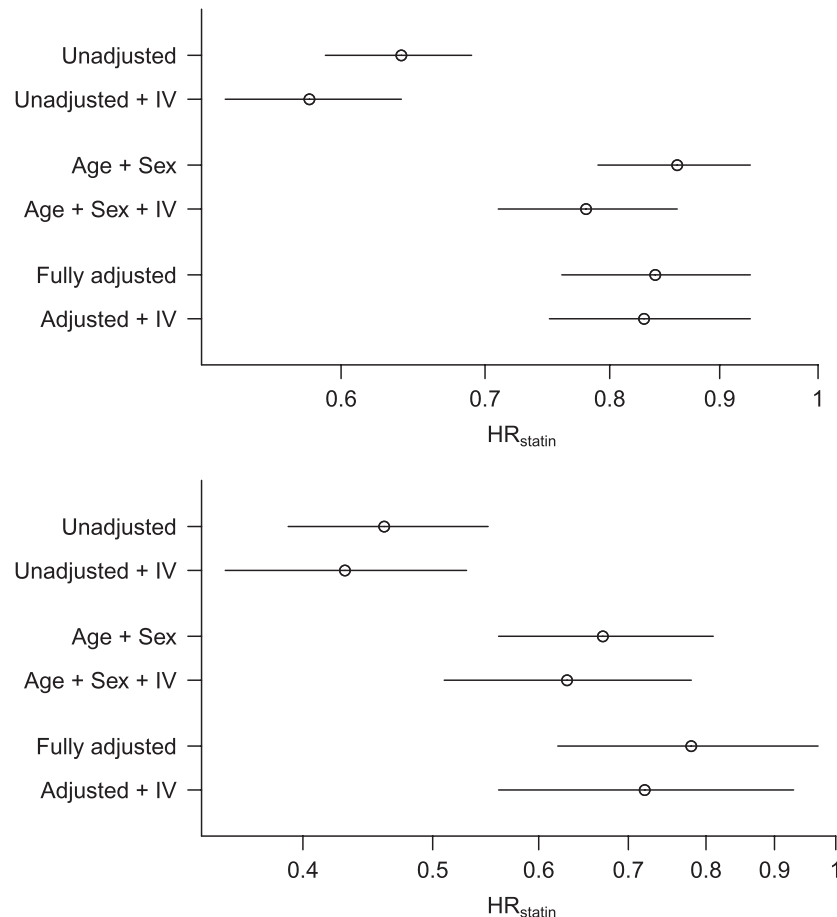
$$E(Y|X = x + 1) - E(Y|X = x) = \beta_2 + \alpha_1\beta_1.$$

This quantity is biased for estimation of  $\beta_2$  owing to confounding from  $U$ , and the bias is equal to  $\alpha_1\beta_1$ . The association between  $X$  and  $Y$  conditional on  $Z$  is given by

$$E(Y|X = x + 1, Z = z) - E(Y|X = x, Z = z) = \beta_2 + \frac{\alpha_1\beta_1}{1 - \alpha_2^2}.$$

The bias of this estimator is  $\alpha_1\beta_1/(1 - \alpha_2^2)$ , which is greater in absolute magnitude than  $\alpha_1\beta_1$  when  $\beta_1$ ,  $\alpha_1$ , and  $\alpha_2$  are all nonzero. If  $\alpha_1$  or  $\beta_1$  is zero, then both estimators are unbiased. If  $\alpha_2 = 0$ , then these biases are equal.

Therefore, in this scenario, conditioning on an IV increases the bias of the exposure effect estimator compared with the unadjusted estimator. This phenomenon can be explained intuitively if we think of partitioning the variation in the exposure variable,  $X$ , into 3 components: the variation explained by  $Z$ , the variation explained by  $U$ , and the unexplained variation. The proportion of the variation explained by  $U$ , along with the association between  $U$  and  $Y$ , determines the magnitude of the unobserved confounding. When we condition on  $Z$ , we effectively remove one source of variation, thereby making the variation explained by  $U$  a larger proportion of the remaining variation in  $X$ . Thus, the residual confounding bias from  $U$  is amplified as a result of



**Figure 2.** Estimated hazard ratios (HRs) for mortality (top) and hip fracture (bottom) in initiators of statin medication versus initiators of glaucoma medication (details in Appendix 2). The adjustment factors used for each estimate, including the potential instrumental variable (IV) glaucoma diagnosis, are shown on the left. The x-axis is presented on the log scale with tick marks unlogged. The approximate expected effects for mortality and hip fracture were hazard ratios of 0.85 and 1.01, respectively (25). Horizontal bars, 95% confidence interval.

conditioning on  $Z$ . This intuition holds whenever there is unobserved confounding and an IV, regardless of the specific assumptions made above. In Appendix 1, we provide an example data set that exhibits bias amplification.

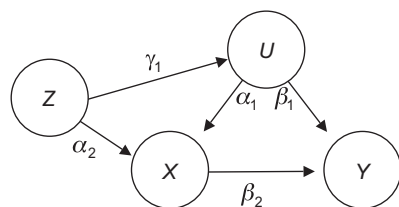
### Empirical example

The example of Patrick et al. (25) provides context for the simulations that follow. In that study, rates of mortality and hip fracture among elderly initiators of statin therapy and glaucoma medications were compared. (The source population and cohort are described in Appendix 2.) Information on demographic characteristics, pretreatment diagnoses, and pretreatment use of health-system services was extracted to define 202 potential confounders. The investigators compared methods of selecting confounders for inclusion in the propensity score model for exposure to statins versus glaucoma drugs. The inclusion of one covariate, prior glaucoma diagnosis, resulted in effect estimates that consistently moved away from the expected effect based on the evidence from randomized controlled trials (see Figure 2).

Glaucoma diagnosis is strongly negatively associated with exposure to statins versus glaucoma drugs (odds ratio = 0.07), but it does not independently predict mortality or hip fracture. Therefore, glaucoma diagnosis appears to be acting as an IV in this example, since its association with exposure is much stronger than its association with outcome, and the observed changes in effect estimates may be a manifestation of bias amplification. Although the analysis of hip fracture is one of the most extreme examples of bias amplification documented in the literature (an increase of 21% in the fully adjusted analysis), so much residual confounding remains that including the IV in the propensity score model does not alter study conclusions. In addition, the strength of the IV-exposure relation in this example makes the IV easy to identify and remove by investigators.

### Monte Carlo simulation studies

Pearl (24) and White and Lu (21) provide formulas for the increases in bias and variance associated with conditioning on an IV or near-IV, but the rescaling of these results to a



**Figure 3.** Causal diagram showing the structure of the simulation studies. Depending on parameter values, the measured covariate  $Z$  may act as a confounder or as an instrumental variable for the exposure-outcome pair  $(X, Y)$ .

given scenario requires considerable computation. Therefore, we performed 2 Monte Carlo simulation studies to obtain quantitative results under a range of epidemiologic scenarios. In the first experiment, we simulated data under an additive model and assumed that the goal of estimation was the risk difference in outcome between levels of exposure. In the second experiment, we simulated data under a multiplicative model and considered the goal of estimation to be the risk ratio for the outcome according to level of exposure. For simplicity and to reflect a common study framework, all variables are binary.

Both simulation studies assumed the same basic causal structure, shown in Figure 3. The true exposure effect and target of estimation is  $\beta_2$ . Note that  $Z$  is not a perfect instrument in Figure 3 as it was in Figure 1 because it is associated with the unmeasured confounder  $U$  through  $\gamma_1$ . However, by varying the value of  $\gamma_1$ , we can explore the impacts of conditioning on  $Z$  when it is a perfect instrument and when it is a near-instrument or confounder. As shown by Pearl (24), bias amplification may result even when the conditioning variable is not a perfect instrument. In addition, we consider relatively large values of  $\gamma_1$  to compare the risks of adjusting for an IV with the benefits of adjusting for a real confounder. The code used to produce and analyze the simulations is available in Web Appendix 1, which appears on the *Journal's* Web site (<http://aje.oxfordjournals.org/>).

### Simulation under additive risk

In each data set, we simulated a binary variable,  $Z$ , with  $\Pr(Z = 1) = 0.5$  and binary variables  $U$ ,  $X$ , and  $Y$ , such that

$$\begin{aligned}\Pr(U = 1 | Z) &= \gamma_0 + \gamma_1 Z, \\ \Pr(X = 1 | U, Z) &= \alpha_0 + \alpha_1 U + \alpha_2 Z, \\ \Pr(Y = 1 | U, X) &= \beta_0 + \beta_1 U + \beta_2 X.\end{aligned}$$

Variables were simulated in the above order so that the risk of outcome would depend directly on  $U$  and  $X$  and indirectly on  $Z$ . The parameters  $\gamma_0$ ,  $\alpha_0$ , and  $\beta_0$  define the baseline prevalence of each variable, and each effect parameter may be interpreted as a risk difference. The values considered for each parameter are listed in Table 1. These values were chosen to provide the widest possible range of scenarios within the  $(0, 1)$  probability bounds for each variable. We considered 2 values for the baseline risk of outcome,  $\beta_0$  in  $\{0.01, 0.2\}$ , corresponding to rare and relatively com-

mon outcomes, respectively. Based on the value of  $\beta_0$ , we constructed a range of possible values for  $\beta_1$ . Within this restriction, we considered all possible combinations of parameter values, resulting in 1,280 unique simulation scenarios. We included only 2 values for the exposure effect,  $\beta_2$ , because bias is invariant to the value of this parameter. We included only positive parameter values to make the illustration of concepts as clear as possible and to avoid repeating scenarios that are symmetric and yield identical results.

For each simulation scenario, we simulated 2,500 data sets of size  $n = 10,000$ . In each data set, we calculated

- the crude risk difference (RD) between  $X$  and  $Y$ ,  $RD_{\text{crude}}$ , and
- the Mantel-Haenszel risk difference (33) between  $X$  and  $Y$  conditional on  $Z$ ,  $RD_{\text{cond}}$ .

Both  $RD_{\text{crude}}$  and  $RD_{\text{cond}}$  are estimators of the exposure effect, and we compared the performance of these two estimators.

### Simulation under multiplicative risk

Using the same binary variable  $Z$  as in the additive study, we simulated binary variables  $U$ ,  $X$ , and  $Y$  such that

$$\begin{aligned}\Pr(U = 1 | Z) &= \gamma_0 \gamma_1^Z, \\ \Pr(X = 1 | U, Z) &= \alpha_0 \alpha_1^U \alpha_2^Z, \\ \Pr(Y = 1 | U, X) &= \beta_0 \beta_1^U \beta_2^X.\end{aligned}$$

Simulating variables in the above order creates data with the causal structure depicted in Figure 3 with associations parameterized as risk ratios. The values considered for each parameter are listed in Table 2. We again considered all possible combinations of parameter values, which resulted in 1,440 unique simulation scenarios. We used multiple values of the true exposure effect,  $\beta_2$ , since bias was no longer invariant to its value.

In each scenario, we simulated 2,500 data sets of size  $n = 10,000$  and calculated

- the crude risk ratio (RR) between  $X$  and  $Y$ ,  $RR_{\text{crude}}$ , and
- the Mantel-Haenszel risk ratio (33) between  $X$  and  $Y$  conditional on  $Z$ ,  $RR_{\text{cond}}$ .

As in the additive simulations, we compared the two estimators of exposure effect,  $RR_{\text{crude}}$  and  $RR_{\text{cond}}$ .

### Evaluation of estimator performance

These simulation studies were designed to compare the performance of estimators for  $\beta_2$  with and without conditioning on  $Z$ . For an estimator of exposure effect  $\hat{\beta}_2$ , we estimated the bias with the equation

$$\text{Bias} = \frac{1}{S} \sum_{s=1}^S \hat{\beta}_2(s) - \beta_2,$$

where  $\hat{\beta}_2(s)$  is the value of  $\hat{\beta}_2$  in the  $s$ th data set and  $S = 2,500$  is the number of simulated data sets. We estimated the standard

**Table 1.** Parameter Values Used in the Additive Simulations<sup>a</sup>

Variable	Baseline Risk	Risk Difference	Corresponding Risk Ratio
<i>U</i>	$\gamma_0 = 0.3$	$\gamma_1$ : 0, 0.006, 0.06, 0.24, 0.6	$\gamma_1$ : 1.0, 1.02, 1.2, 1.8, 3.0
<i>X</i>	$\alpha_0 = 0.3$	$\alpha_1$ : 0, 0.06, 0.18, 0.33	$\alpha_1$ : 1.0, 1.2, 1.6, 2.1
		$\alpha_2$ : 0, 0.06, 0.18, 0.33	$\alpha_2$ : 1.0, 1.2, 1.6, 2.1
<i>Y</i>	$\beta_0 = 0.2$	$\beta_1$ : 0, 0.08, 0.36, 0.5	$\beta_1$ : 1.0, 1.4, 2.8, 3.5
		$\beta_2$ : 0, 0.2	$\beta_2$ : 1.0, 2.0
	$\beta_0 = 0.01$	$\beta_1$ : 0, 0.004, 0.018, 0.5	$\beta_1$ : 1.0, 1.4, 2.8, 51
		$\beta_2$ : 0, 0.2	$\beta_2$ : 1.0, 21.0

<sup>a</sup> The value of  $\beta_0$  determines the set of potential values for  $\beta_1$  and  $\beta_2$ . Within that restriction, all possible combinations of parameter values were considered. The corresponding risk ratios are calculated on the basis of the baseline prevalence of each variable and will vary depending on the values of other variables.

error of  $\hat{\beta}_2$  using the square root of the sample variance of  $\hat{\beta}_2(s)$  across simulated data sets. We calculated the bias and variance of the exposure effect estimators separately in each simulation scenario.

## RESULTS

### Additive simulation

Figure 4 shows the performance of  $RD_{crude}$  on the *x*-axis versus that of  $RD_{cond}$  on the *y*-axis. The left panel displays the biases of both estimators, and the right panel shows the standard errors. Results are shown for all simulation scenarios with  $\beta_0 = 0.2$ ,  $\beta_2 = 0$ , and 3 values of  $\gamma_1$ . All values of  $\beta_1$ ,  $\alpha_1$ , and  $\alpha_2$  are shown; the values of  $\alpha_1$  and  $\beta_1$  are not differentiated (but they may be inferred from the amount of crude bias for a given scenario). Results for other values of  $\beta_0$ ,  $\beta_2$ , and  $\gamma_1$  are similar to the results shown here and are available in Web Appendix 2 (Web Figures 1–4). In each plot, the solid diagonal marks equality. A point on the line indicates a simulation scenario where the bias or standard error is invariant to conditioning on *Z*; scenarios where the bias or standard error is increased or decreased by conditioning on *Z* are represented by points above or below the line, respectively.

In the top row of plots in Figure 4,  $\gamma_1$  equals 0, indicating that *Z* is simulated to be a perfect instrument for the exposure-

outcome pair (*X*, *Y*). Therefore, the bias in  $RD_{crude}$  is due to unobserved confounding from *U*. In general, conditioning on the instrument, *Z*, results in an estimator of exposure effect that is more biased than the crude estimator. In addition, the standard error of  $RD_{cond}$  is often larger than the standard error of  $RD_{crude}$ . The magnitude of these increases depends on the value of  $\alpha_2$ . When *Z* is a strong instrument ( $\alpha_2 = 0.33$ ), the increases in bias and standard error due to conditioning on *Z* are largest; when *Z* is a weak instrument ( $\alpha_2$  in {0.06, 0.18}), the increases are negligible; when *Z* has no association with exposure ( $\alpha_2 = 0$ ), there is no increase in either bias or standard error.

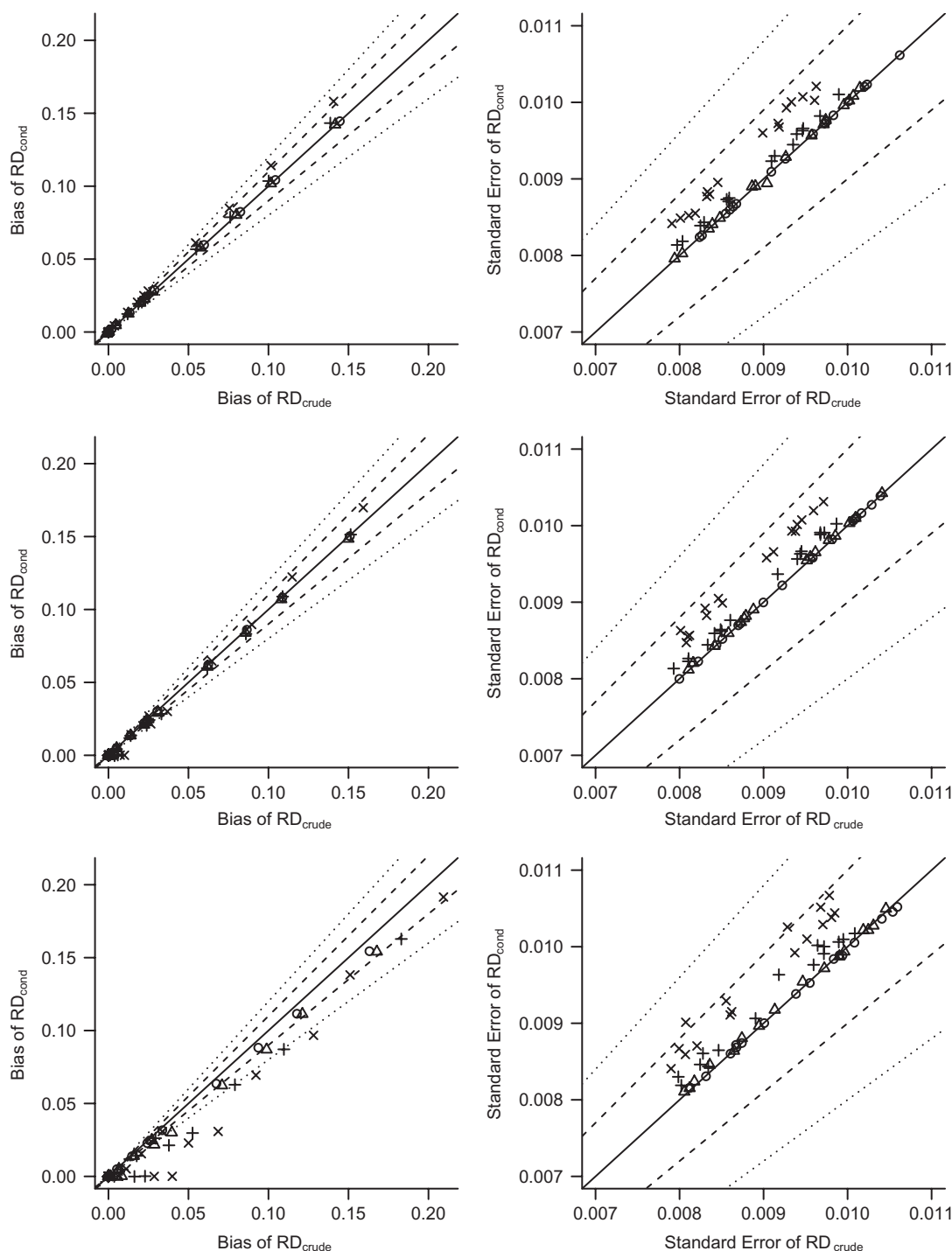
In the center row of Figure 4,  $\gamma_1$  equals 0.06, indicating that *Z* is not a perfect instrument because *Z* is associated with *Y* through the unobserved confounder, *U*. However, we may consider *Z* to be a near-instrument (or near-confounder), since its association with *U* is relatively weak. In these scenarios, conditioning on *Z* tended to result in increased bias in simulation scenarios with the largest crude bias and decreased bias in simulation scenarios with smaller crude bias. In the former case, the unobserved confounding due to *U* overwhelms the relatively small amount of confounding due to *Z*. In the latter case, the confounding due to *U* is smaller, and *Z* accounts for more of the overall confounding bias of exposure effect. The effect on standard error was similar to that observed in the top row (where *Z* is a perfect IV).

**Table 2.** Parameter Values Used in the Multiplicative Simulations<sup>a</sup>

Variable	Baseline Risk	Risk Ratio	Corresponding Risk Difference
<i>U</i>	$\gamma_0 = 0.3$	$\gamma_1$ : 1, 1.02, 1.2, 1.8, 3	$\gamma_1$ : 0, 0.06, 0.06, 0.24, 0.6
<i>X</i>	$\alpha_0 = 0.3$	$\alpha_1$ : 1, 1.1, 1.3, 1.8	$\alpha_1$ : 0, 0.03, 0.09, 0.24
		$\alpha_2$ : 1, 1.1, 1.3, 1.8	$\alpha_2$ : 0, 0.03, 0.09, 0.24
<i>Y</i>	$\beta_0 = 0.2$	$\beta_1$ : 1, 1.2, 2.2	$\beta_1$ : 0, 0.04, 0.24
		$\beta_2$ : 1, 1.2, 2.2	$\beta_2$ : 0, 0.04, 0.24
	$\beta_0 = 0.01$	$\beta_1$ : 1, 2.2, 8.0	$\beta_1$ : 0, 0.012, 0.07
		$\beta_2$ : 1, 2.2, 8.0	$\beta_2$ : 0, 0.012, 0.07

<sup>a</sup> The value of  $\beta_0$  determines the set of potential values for  $\beta_1$  and  $\beta_2$ . Within that restriction, all possible combinations of parameter values were considered. The corresponding risk differences are calculated on the basis of the baseline risk of each variable and will vary depending on the values of other variables.

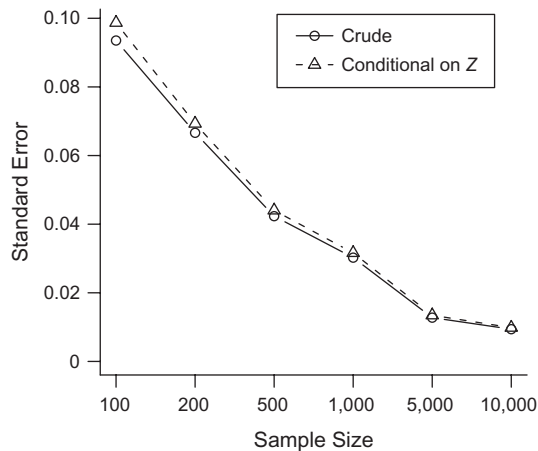




**Figure 4.** Bias (left panels) and standard error (right panels) of risk difference (RD) estimators with and without conditioning on Z. Each point represents one simulation scenario in the additive simulations with  $\gamma_1 = 0$  (upper sections),  $\gamma_1 = 0.06$  (middle sections), or  $\gamma_1 = 0.24$  (lower sections). The symbols identify values of  $\alpha_2$  ( $\circ$ , zero;  $\triangle$ , 0.06;  $+$ , 0.18;  $\times$ , 0.33). The solid diagonal line marks equality. Dashed lines mark the threshold for a 10% increase or decrease, and dotted lines mark a 20% increase or decrease.

In the bottom row of Figure 4,  $\gamma_1$  equals 0.24, indicating that Z is a confounder in these scenarios. When we condition on the confounder, the bias is always equivalent or decreased,

but the standard error may increase or decrease. As before, the magnitude of the increase in standard error is determined by the value of  $\alpha_2$ , with the largest increases occurring when



**Figure 5.** Standard error of exposure effect estimators obtained with and without conditioning on  $Z$  under a range of study sizes.

$\alpha_2$  equals 0.33. Furthermore, when  $\alpha_2$  equals 0,  $Z$  has no direct association with exposure, but conditioning on  $Z$  still reduces the bias.

Across all of the additive simulation scenarios defined in Table 1, the largest absolute increase in bias due to conditioning on  $Z$  was an increase of 0.018 on a crude bias of 0.141. This scenario had the highest value considered for each of  $\alpha_1$ ,  $\beta_1$ , and  $\alpha_2$  and  $\gamma_1 = 0$ . (Equal biases were found across values of  $\beta_0$  and  $\beta_2$ .) The largest observed increase in standard error due to conditioning on  $Z$  was an increase of 0.003 on a crude standard error of 0.009. This scenario had the highest value considered for all parameters.

Because  $\alpha_2$  is shown to be the most important parameter in determining the magnitude of the increases in bias and variance when conditioning on  $Z$ , we further considered a scenario with a larger value for  $\alpha_2$ . In the case of a binary exposure, the value of  $\alpha_2$  is constrained by the (0, 1) bounds on probability of exposure. Therefore, in order to increase  $\alpha_2$ , we reduced the baseline prevalence of exposure ( $\alpha_0 = 0.1$ ) and chose the other parameter values as follows:  $\gamma_0 = 0.3$ ,  $\gamma_1 = 0$ ,  $\alpha_2 = 0.6$ ,  $\beta_0 = 0.2$ ,  $\beta_1 = 0.5$ , and  $\beta_2 = 0$ . Simulating with these values yielded biases of 0.101 and 0.158 for  $RD_{\text{crude}}$  and  $RD_{\text{cond}}$ , respectively, representing a 56% increase in bias. The standard errors of  $RD_{\text{crude}}$  and  $RD_{\text{cond}}$  were 0.01 and 0.012, respectively, representing a 20% increase in standard error.

We further repeated one simulation scenario under varying study sizes to explore the bias-variance trade-off as study size is reduced. In particular, we use the scenario reported above with the largest absolute increase in bias due to conditioning on  $Z$  from Table 1. Figure 5 displays the standard error of  $RD_{\text{crude}}$  and  $RD_{\text{cond}}$  under a range of study sizes. The standard error increases rapidly as the study size decreases, and the increase in standard error attributable to conditioning on  $Z$  is negligible compared with the impact of study size. In addition, even at the smallest study size considered ( $n = 100$ ), the standard errors of both  $RD_{\text{crude}}$  and  $RD_{\text{cond}}$  are smaller than the bias in this scenario.

## Multiplicative simulation

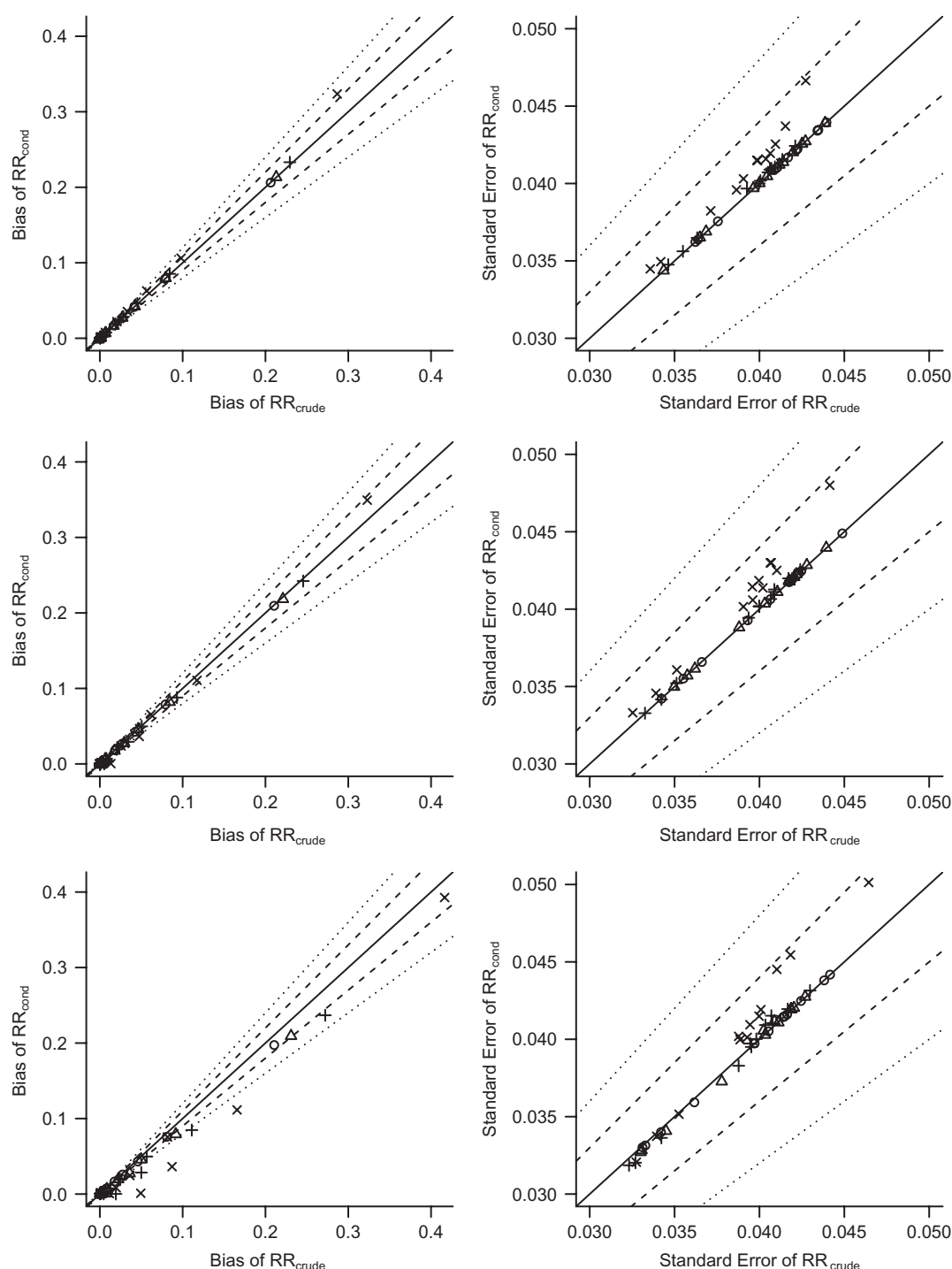
Figure 6 shows the bias (left) and standard error (right) of  $RR_{\text{crude}}$  on the  $x$ -axis versus that of  $RR_{\text{cond}}$  on the  $y$ -axis. As in Figure 4, the  $y = x$  line is provided. Results are displayed for all simulation scenarios with  $\beta_0 = 0.2$ ,  $\beta_2 = 1$ , and 3 values of  $\gamma_1$ . Results for other values of  $\beta_0$ ,  $\beta_2$ , and  $\gamma_1$  are similar to the results shown here and are available in Web Appendix 2 (Web Figures 5–10).

In the multiplicative simulations, associations are parameterized as risk ratios, so the 3 values of  $\gamma_1$  shown in Figure 6 indicate that the variable  $Z$  is simulated to be a perfect instrument, a near-instrument (or near-confounder), and a confounder, respectively, for the exposure-outcome pair ( $X$ ,  $Y$ ). Results are similar to the results from the additive simulations. In the presence of unobserved confounding, conditioning on a true instrument increases the bias and standard error in exposure effect estimation, and this increase tends to be larger when the instrument is strong ( $\alpha_2 = 1.8$ ) and when the crude bias or standard error is large. In the scenarios with no confounding bias from  $U$  ( $\alpha_1 = 1$  or  $\beta_1 = 1$ ), conditioning on  $Z$  does not create bias. Conditioning on a near-instrument tends to result in increased bias when the crude bias is large and decreased bias when the crude bias is relatively small. When  $Z$  is a confounder, bias generally decreases as a result of conditioning on  $Z$ , but standard error may increase or decrease.

The largest absolute bias increase for any scenario was an increase of 1.636 on a crude bias of 7.773, achieved when  $\gamma_1 = 1$ ,  $\beta_0 = 0.01$ , and the parameters  $\alpha_1$ ,  $\alpha_2$ ,  $\beta_1$ , and  $\beta_2$  are maximized. The same scenario results in the largest increase in standard error across all multiplicative simulation scenarios: an increase of 0.25 on a crude standard error of 1.676. The scale of both bias and standard error is larger in the multiplicative simulations than in the additive simulations, but bias remains the primary source of error.

## DISCUSSION

Our simulation studies showed that estimating an exposure effect conditional on a perfect instrument can increase the bias and standard error of the exposure effect estimate, but these increases were generally small. In particular, when residual confounding was small, the increase in bias and variance due to conditioning on an IV was essentially negligible. When the residual confounding bias was large, the increase in estimation error due to conditioning on an IV represented only a small fraction of the overall error in most scenarios. In addition, increases in bias and standard error were observed when conditioning on a variable that was strongly associated with exposure and weakly associated with outcome. These increases were always smaller than the increases observed when adjusting for a perfect IV with equivalent association with exposure. As expected, the effects of conditioning on an IV or near-IV were reduced with diminishing strength of the unmeasured confounding and diminishing strength of the IV association with exposure. These results are consistent with past theoretical and simulation findings (18, 20–24, 34).



**Figure 6.** Bias (left panels) and standard error (right panels) of risk ratio (RR) estimators with and without conditioning on Z. Each point represents one simulation scenario in the multiplicative simulations with  $\gamma_1 = 1$  (upper sections),  $\gamma_1 = 1.2$  (middle sections), or  $\gamma_1 = 1.8$  (lower sections). The symbols identify values of  $\alpha_2$  (○, 1.0; △, 1.1; +, 1.3; ×, 1.8). The solid diagonal line marks equality. Dashed lines mark the threshold for a 10% increase or decrease, and dotted lines mark a 20% increase or decrease.

These results have clear implications for epidemiologic practice. First, variables that are known to be instruments should not be conditioned upon. The belief that balancing

all preexposure covariates, as in randomized studies, can do no harm does not hold in nonexperimental studies because there may exist unobserved factors that cannot be balanced.



Contrary to the current practice of selecting the best predictors of exposure, a very strong association with exposure may be indicative of an IV or near-IV that should be excluded, as shown in the example study. Second, ordering variables based on the magnitude of their association with outcome could provide a reasonable approach to selecting covariates for conditioning, as recommended by Hill (35) and implemented in a high-dimensional propensity score algorithm (31) and in Bayesian propensity scores (36). Although IVs may be associated with outcome in the presence of unmeasured confounding, covariates with relatively strong associations with outcome are unlikely to be IVs. Finally, within the context of scenarios considered in the simulation studies, inadvertently including an IV in the set of conditioning variables does not appear to pose a major threat to the validity of exposure effect estimates. In most scenarios, the need to control residual confounding greatly outweighed bias amplification caused by adjusting for an IV. This threat can be further reduced if strong predictors of exposure are carefully considered before being used in adjustment.

Although we were able to deduce consistent trends across simulation scenarios, specific findings are dependent on the specification of the data-generating process and the parameter values considered. In particular, it is clear that the magnitude of the increase in bias is limited only by the extent to which the IV determines exposure. In the case of a binary exposure, this parameter is constrained by the baseline prevalence of exposure and the effects of other factors that determine exposure. When analyzing a continuous exposure, no such constraints exist, and the IV association with exposure may be larger. In addition, in cases of a known IV (e.g., randomized assignment to exposure), the association between the IV and exposure may be stronger. In our simulation studies, the parameter values were chosen to represent the range of associations most likely to be encountered in epidemiologic studies with a binary exposure and a binary covariate that is not known to be an IV. Within this range, the maximum increase in bias (over the crude bias) observed in any scenario was approximately 20%. When we further considered a scenario with a stronger association between the IV and exposure, we observed a 56% increase in bias. However, achieving this magnitude of bias increase required both extremely large unmeasured confounding and a very strong instrument. On the other hand, when conditioning on a confounder, a 50% or greater decrease in bias was relatively easy to achieve and did not require such an extreme scenario.

## ACKNOWLEDGMENTS

Author affiliations: Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts (Jessica A. Myers, Jeremy A. Rassen, Joshua J. Gagne, Krista F. Huybrechts, Sebastian Schneeweiss, Kenneth J. Rothman, Robert J. Glynn); RTI Health Solutions, Research Triangle Park, North Carolina (Kenneth J. Rothman); and Department of Biostatistics and Epidemiology, School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania (Marshall M. Joffe).

This project was funded by the Food and Drug Administration through the Mini-Sentinel Coordinating Center and by grant RO1-LM010213 from the National Library of Medicine.

The authors acknowledge the members of the Food and Drug Administration Mini-Sentinel Signal Evaluation Working Group for their contributions to this paper.

Conflict of interest: none declared.

## REFERENCES

1. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*. 1968;24(2):295–313.
2. Billewicz WZ. The efficiency of matched samples: an empirical investigation. *Biometrics*. 1965;21(3):623–644.
3. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41–55.
4. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc*. 1984;79(387):516–524.
5. D'Agostino RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med*. 1998;17(19):2265–2281.
6. Rosenbaum PR. *Observational Studies*. 2nd ed. New York, NY: Springer Verlag, Publishers; 2002.
7. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat Med*. 2007;26(1):20–36.
8. Rubin DB. Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? *Stat Med*. 2009;28(9):1420–1423.
9. Weitzen S, Lapane KL, Toledano AY, et al. Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiol Drug Saf*. 2004;13(12):841–853.
10. Shrier I, Re: The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials [letter]. *Stat Med*. 2008;27(14):2740–2741.
11. Pearl J. *Causality: Models, Reasoning, and Inference*. 2nd ed. New York, NY: Cambridge University Press; 2009.
12. Sjölander A. Propensity scores and M-structures. *Stat Med*. 2009;28(9):1416–1420.
13. Martens EP, Pestman WR, de Boer A, et al. Instrumental variables: application and limitations. *Epidemiology*. 2006;17(3):260–267.
14. Glymour MM. Natural experiments and instrumental variable analyses in social epidemiology. In: Oakes JM, Kaufman JS, eds. *Methods in Social Epidemiology*. San Francisco, CA: John Wiley & Sons, Inc; 2006:429–468.
15. Hernán MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? *Epidemiology*. 2006;17(4):360–372.
16. Grootendorst P. A review of instrumental variables estimation of treatment effects in the applied health sciences. *Health Serv Outcomes Res Methodol*. 2007;7(3):159–179.
17. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med*. 1997;127(8):757–763.
18. Brookhart MA, Schneeweiss S, Rothman KJ, et al. Variable selection for propensity score models. *Am J Epidemiol*. 2006;163(12):1149–1156.
19. Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance

- measured variables between treated and untreated subjects: a Monte Carlo study. *Stat Med.* 2007;26(4):734–753.
20. Hahn J. Functional restriction and efficiency in causal inference. *Rev Econ Stat.* 2004;86(1):73–76.
  21. White H, Lu X. Causal diagrams for treatment effect estimation with application to efficient covariate selection. *Rev Econ Stat.* In press.
  22. Bhattacharya J, Vogt WB. *Do Instrumental Variables Belong in Propensity Scores?* (NBER Technical Working Paper no. 343). Cambridge, MA: National Bureau of Economic Research; 2007.
  23. Wooldridge J. *Should Instrumental Variables Be Used As Matching Variables?* East Lansing, MI: Michigan State University; 2009. (<https://www.msu.edu/~ec/faculty/wooldridge/current%20research/treat1r6.pdf>). (Accessed May 1, 2011).
  24. Pearl J. On a class of bias-amplifying variables that endanger effect estimates. In: Grünwald P, Spirtes P, eds. *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI 2010)*. Corvallis, OR: Association for Uncertainty in Artificial Intelligence; 2010:425–432.
  25. Patrick AR, Schneeweiss S, Brookhart MA, et al. The implications of propensity score variable selection strategies in pharmacoepidemiology: an empirical illustration. *Pharmacoepidemiol Drug Saf.* 2011;20(6):551–559.
  26. Stukel TA, Fisher ES, Wennberg DE, et al. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. *JAMA.* 2007;297(3):278–285.
  27. Novikov I, Kalter-Leibovici O. Analytic approaches to observational studies with treatment selection bias [letter]. *JAMA.* 2007;297(19):2077.
  28. D'Agostino RB Jr, RB D'Agostino Sr. Estimating treatment effects using observational data. *JAMA.* 2007;297(3):314–316.
  29. Stukel TA, Fisher ES, Wennberg DE. Analytic approaches to observational studies with treatment selection bias—reply [letter]. *JAMA.* 2007;297(19):2078.
  30. Stukel TA, Fisher ES, Wennberg DE. Using observational data to estimate treatment effects. *JAMA.* 2007;297(19):2078–2079.
  31. Schneeweiss S, Rassen JA, Glynn RJ, et al. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology.* 2009;20(4):512–522.
  32. Brookhart MA, Stürmer T, Glynn RJ, et al. Confounding control in healthcare database research: challenges and potential approaches. *Med Care.* 2010;48(6 suppl):S114–S120.
  33. Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008.
  34. Rubin DB, Thomas N. Matching using estimated propensity scores: relating theory to practice. *Biometrics.* 1996;52(1):249–264.
  35. Hill J. Discussion of research using propensity-score matching: comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003' by Peter Austin. *Statistics in Medicine. Stat Med.* 2008;27(12):2055–2061.
  36. McCandless LC, Gustafson P, Austin PC. Bayesian propensity score analysis for observational data. *Stat Med.* 2009;28(1):94–112.

## APPENDIX 1

### Example Data Set

We present an example data set from the multiplicative simulation study, where amplification of bias and standard error were relatively large. These data were simulated under the following true parameter values:  $\beta_0 = 0.01$ ,  $\beta_1 = 8$ ,  $\beta_2 = 8$  (the true exposure effect),  $\alpha_0 = 0.3$ ,  $\alpha_1 = 1.8$ ,  $\alpha_2 = 1.8$ ,  $\gamma_0 = 0.3$ , and  $\gamma_1 = 1$ . The simulated data for one set of 10,000 patients is given in Appendix Table 1. The crude estimate of exposure effect from these data is  $RR_{\text{crude}} = 15.52$ . Thus, the bias of  $RR_{\text{crude}}$  is 7.52 ( $15.52 - 8 = 7.52$ ). The estimate of exposure effect conditional on Z is  $RR_{\text{cond}} = 17.07$ , and the bias of  $RR_{\text{cond}}$  is 9.07 ( $17.07 - 8 = 9.07$ ). Note that the same mechanism that results in bias amplification also results in estimates of exposure effect that are heterogeneous across strata of Z (risk ratios of 12.7 when  $Z = 0$  and 26.8 when  $Z = 1$ ).

**Appendix Table 1.** One Simulated Data Set From the Multiplicative Simulations

	Z = 0		Z = 1	
	Y = 1	Y = 0	Y = 1	Y = 0
X = 1	603	1,258	1,084	2,263
X = 0	80	3,059	20	1,633

## APPENDIX 2

### Empirical Example

The data in the empirical example come from the investigation described by Patrick et al. (25). That cohort study included patients initiating the use of statins and glaucoma medications among Medicare beneficiaries aged 65 years or older who were enrolled in the Pharmaceutical Assistance Contract for the Elderly (PACE) program provided by the state of Pennsylvania. Enrollees in PACE were eligible for inclusion in the study population if they had filled a prescription for any statin or glaucoma drug between January 1, 1996, and December 31, 2002, and demonstrated continuous use of the health-care system.

Initiation of drug therapy was defined as an eligible beneficiary's filling at least 1 prescription for a medication of interest between January 1, 1996, and December 31, 2002, but not using one during the 18 months prior to the index date. The index date was the first date on which a prescription for a statin or glaucoma drug was filled. Follow-up was continued for 1 year after the initiation of therapy. Covariates were defined on the basis of enrollment information (age, sex, race) and claims made during the year before the index date.